An asymptotic test of optimality conditions in multiresponse simulation optimization

Angun, M.E.; Kleijnen, Jack P.C.

Published in:
INFORMS Journal on Computing

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

INFORMS Journal on Computing
Vol. 24, No. 1, Winter 2012, pp. 53–65
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.1100.0438

An Asymptotic Test of Optimality Conditions in Multiresponse Simulation Optimization

Ebru Angün
Department of Industrial Engineering, Galatasaray University, Ortaköy, 34357 İstanbul, Turkey, eangun@gsu.edu.tr
Jack Kleijnen
Department of Information Management/CentER, Tilburg University, 5000 LE Tilburg, The Netherlands, kleijnen@uvt.nl

This paper derives a novel, asymptotic statistical test of the Karush–Kuhn–Tucker first-order necessary optimality conditions in random simulation models with multiple responses. This test combines a simple form of the delta method and a generalized version of Wald's statistic. The test is applied to both a toy problem and an (s, S) inventory-optimization problem with a service-level constraint; its numerical results are encouraging.

Key words: simulation, design of experiments; simulation, statistical analysis; black-box simulation-optimization; white-box simulation-optimization
History: Accepted by Marvin Nakayama, Area Editor for Simulation; received March 2010; revised September 2010; accepted September 2010. Published online in *Articles in Advance* February 2, 2011.

1. Introduction

In this paper, we present a novel heuristic procedure to test the optimality of a given input vector for a random simulation model with multiple responses, where one of these responses is selected as the objective whereas the remaining responses should satisfy prespecified threshold values. More specifically, we derive a statistical stopping rule that assumes many replicates; the term "replicates" implies that a particular input vector is simulated several times, using nonoverlapping streams of random variates. Our stopping rule tests the well-known Karush–Kuhn–Tucker (KKT) first-order necessary optimality conditions at a feasible point—originally derived for deterministic optimization (see, for example, Gill et al. 2000, p. 81). In the rest of this paper, these conditions will be abbreviated to the KKT conditions.

There are many methods for optimizing simulated systems (see, for example, the survey paper by Fu 2002 or the monograph by Spall 2003). Many methods ignore the fact that in practice simulation models generate multiple responses per input vector. For example, an academic (s, S) inventory simulation—with reorder level s and order-up-to level S—has as a response the expected (or mean) sum of the inventory-carrying, ordering, and out-of-stock costs; a practical simulation model, however, typically has two responses, namely, the sum of the average inventory-carrying and ordering costs—which is to be minimized—and the service probability (also called the fill-rate)—which must satisfy a prespecified lower bound (say, 95%). In this paper, we minimize a stochastic objective function under one or more constraints on the remaining stochastic responses.

Some of these optimization methods treat the simulation model as a black box; i.e., they assume that no gradient information is available. Examples are the many metaheuristics (ant colony optimization, genetic and evolutionary algorithms, scatter search, simulated annealing, tabu search), simultaneous perturbation stochastic approximation, and response surface methodology (RSM). Other methods treat the simulation as a white box; i.e., they estimate the gradients from a single simulation run—best known are perturbation analysis and the score function (or likelihood ratio) method. Our procedure can be combined with any method that estimates gradients (either from a single run or from several runs), provided the gradient estimators satisfy certain technical conditions—which we shall introduce in §2.

There are a few related publications. Bettonvil et al. (2009) address the same problem as we do, but those authors assume that the number of replicates is small (so asymptotic statistical properties do not hold, and bootstrapping is used instead). Karaesmen and van Ryzin (2004) check the KKT conditions for a single stochastic response (they estimate the gradient through the score function method). Higle and Sen (1991) provide termination criteria that are based on statistical verification of generalized KKT conditions and conditions based on Lagrange duality using

a bootstrap procedure in a stochastic programming context.

Our procedure provides a stopping criterion for iterative (heuristic) simulation-based constrained or unconstrained optimization. The two important statistical components of our stopping rule are as follows:

1. The delta method, which shows that under specific conditions, nonlinear statistics are asymptotically multivariate normally distributed; see, for example, Rubinstein and Shapiro (1993, p. 259).

2. The generalized Wald statistic presented in Kodde and Palm (1986), which enables the testing of composite hypotheses.

However, we do not incorporate our rule into an optimization algorithm because that would require answering many other questions, such as when to use the rule so that the overall Type I (also known as type α) error rate of the optimization process is bounded. We leave this implementation for future research.

In hypothesis testing, the users typically determine an upper bound for the Type I error probability. Theoretically, it is also possible to constrain the Type II error (or β-error) by simulating enough replicates; in practice, however, it is not possible to determine how many replicates are enough. Nevertheless, it is desired that the Type II error probability decreases as the alternative hypothesis deviates more from the null hypothesis (i.e., the power function of the test increases). Stage 1 of our test, which uses a *Student t-test*, has the desired behavior under asymptotic normality. Therefore, we empirically investigate the behavior of Stages 2 and 3 of our test, which use the *bootstrap's percentile method* and the generalized form of *Wald's statistic*, respectively, and we find that the power function for Stage 3 indeed has the desired shape in a toy problem.

The remainder of this paper is organized as follows. In §2, we formalize our problem including some statistical issues. In §3, we formalize the proposed stopping rule. In §4, we evaluate this stopping rule through a toy problem and the (s, S) inventory optimization problem with a service-level constraint, which is originally investigated by Bashyam and Fu (1998). In §5, we present conclusions. In the online supplement (available at http://joc.journal.informs.org/), we present the appendix with technical details and additional numerical results.

2. Problem Formulation and Statistical Techniques

In this section, we present the mathematical formulation of our problem and explain the standard statistical techniques for the local approximation of the unknown functions in our problem. The only nonstandard technique is introduced in §2.1, which is a procedure for determining the size of the local experimental area.

We assume that we can model our problem mathematically as follows:

$$\begin{aligned} &\text{minimize } E_{\boldsymbol{\omega}}[f_0(\mathbf{d}, \boldsymbol{\omega})] \\ &\text{subject to } E_{\boldsymbol{\omega}}[f_j(\mathbf{d}, \boldsymbol{\omega})] \geq a_j \quad \text{for } j = 1, \ldots, r-1, \end{aligned} \tag{1}$$

where $E_{\boldsymbol{\omega}}$ is the expectation operator with respect to the simulation's seed vector $\boldsymbol{\omega}$, the f_i $(i = 0, \ldots, r-1)$ are r random responses that are computed through simulation, $\mathbf{d} = (d_1, \ldots, d_k)^T$ is the $k \times 1$ input vector, and a_j is the deterministic right-hand-side value for the jth constraint. An example of (1) is the minimization of the expected inventory-carrying and ordering costs such that a prespecified customer service level is satisfied; obviously, r then equals 2.

We now explain how to approximate the simulation responses in (1) locally through second-order polynomials in $\mathbf{d}$. We can estimate the unknown coefficients of these polynomials and the covariance matrices of the estimated coefficients through perturbation analysis or score function methods (i.e., we treat the simulation as a white box), the simultaneous perturbation method, or linear regression; the latter two methods treat the simulation as a black box. There are two technical conditions that the sequence of gradient estimators must satisfy (i) the sequence converges in probability to the true gradient, and (ii) the normalized sequence converges in distribution to multivariate normal. We need these two conditions and the covariance matrices when we apply the delta method (e.g., Rubinstein and Shapiro 1993, p. 259) and the generalized Wald statistic (Kodde and Palm 1986) to derive our stopping rule. Actually, in this paper we estimate the gradients and their covariance matrices using linear regression as follows.

We locally approximate the r simulation responses in (1) by r second-order polynomial regression metamodels:

$$g_i(\mathbf{d}, \boldsymbol{\omega}) = \beta_{i;0} + \sum_{t=1}^{k} \beta_{i;t} d_t + \sum_{t=1}^{k} \sum_{t' \geq t}^{k} \beta_{i;t;t'} d_t d_{t'} + \epsilon_i(\mathbf{d}, \boldsymbol{\omega}) \quad \text{for } i = 0, \ldots, r-1, \tag{2}$$

where $\boldsymbol{\beta}_i = (\beta_{i;0}, \beta_{i;1}, \ldots, \beta_{i;k}, \beta_{i;1;1}, \beta_{i;1;2}, \ldots, \beta_{i;k-1;k}, \beta_{i;k;k})^T$ is the $q \times 1$ vector of unknown regression coefficients with $q = 1 + 2k + k(k-1)/2$, and $\boldsymbol{\epsilon}_i$ is the additive error that results from both inherent randomness in stochastic simulation caused by $\boldsymbol{\omega}$ in (1) and possible lack of fit of the second-order polynomial approximation.

To fit the second-order polynomials in (2), simulation practitioners typically use a central composite

design (CCD), which is defined as follows (also see Kleijnen 2008). One part of a CCD consists of a two-level factorial design that may be fractional—provided this fractional has a resolution at least V—because a resolution V (R_V) design gives unbiased ordinary least squares (OLS) estimators of $\beta_{i;0}$, $\beta_{i;t}$s, and $\beta_{i;t;t'}$s with $t \neq t'$—provided all other effects are negligible. Furthermore, to estimate $\beta_{i;t;t'}$s with $t = t'$, a CCD augments a R_V design by (i) the central input vector (input combination) and (ii) $2k$ "axial" input vectors, which increase and decrease the central input vector one at a time (also see §2.1). Therefore, a CCD requires the explicit determination of $n = n_V + 2k + 1$ input vectors in the local experimental area where n_V denotes the number of input vectors in the R_V design; we explain this procedure in §2.1.

Classic linear regression assumes a single response ($r = 1$) and white noise ϵ_i in (2); that is, ϵ_i is independent, identically distributed (IID) normal with mean $\mu_{\epsilon_i} = 0$ and a homoscedastic (constant) variance $\sigma^2_{\epsilon_i}$ across the n input vectors. The plausibility of the white-noise assumption is discussed in Kleijnen (2008, p. 20). We assume that the white-noise assumption holds per response i within the local experimental area. In the numerical examples, we investigate the effects of departure from this assumption; in particular, we consider cases in which the variance changes with the input vector and $\mu_{\epsilon_i} \neq 0$. We also emphasize that the variance is homoscedastic only within the local experimental area, so it must be reestimated whenever this area changes.

When the white-noise assumption is satisfied, OLS gives the best linear unbiased estimator (BLUE) of the $\boldsymbol{\beta}_i$ where "best" means minimum variance (Theil 1971, p. 119). In our problem, however, we have r responses that are correlated (because they are generated by the same seed vector). Moreover, the variances differ among the r simulation responses. Hence, generalized least squares (GLS) is needed to find the BLUE of $\boldsymbol{\beta}_i$. Rao (1967), however, proves that if the same design is used for all r responses and if each individual response satisfies the white-noise assumption, then the GLS estimator of the vector $(\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_{r-1}^T)^T$ reduces to the OLS estimators of the individual coefficient vectors $\boldsymbol{\beta}_i$:

$$\hat{\boldsymbol{\beta}}_i^N(\mathbf{d}, \boldsymbol{\omega}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{f}}_i(\mathbf{d}, \boldsymbol{\omega}) \quad \text{for } i = 0, \ldots, r-1, \tag{3}$$

where $\mathbf{X}$ is the $N \times q$ matrix of explanatory variables, $N = m_1 + \cdots + m_l + \cdots + m_n$ is the total number of runs in the local experimental area, m_l is the number of replicates made at input vector $\mathbf{d}_l (l = 1, \ldots, n)$, and $\hat{\mathbf{f}}_i(\mathbf{d}, \boldsymbol{\omega}) = (\hat{\mathbf{f}}_i(\mathbf{d}_1, \boldsymbol{\omega}_1)^T, \ldots, \hat{\mathbf{f}}_i(\mathbf{d}_l, \boldsymbol{\omega}_l)^T, \ldots, \hat{\mathbf{f}}_i(\mathbf{d}_n, \boldsymbol{\omega}_n)^T)^T$ is the $N \times 1$ vector of IID simulated responses i with $\hat{\mathbf{f}}_i(\mathbf{d}_l, \boldsymbol{\omega}_l)$ being the m_l IID replicates simulated at $\mathbf{d}_l$ with the seed vector $\boldsymbol{\omega}_l$.

The covariance matrix of $\hat{\boldsymbol{\beta}}_i^N$ in (3) is

$$\widehat{\boldsymbol{\Psi}}_i^N(\mathbf{d}, \boldsymbol{\omega}) = \hat{\sigma}_{i;i}^N(\mathbf{d}, \boldsymbol{\omega})(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{for } i = 0, \ldots, r-1, \tag{4}$$

where $\hat{\sigma}_{i;i}^N$ is an estimator of the variance $\sigma_{i;i}$ of response i. We estimate these $\sigma_{i;i}$ through the standard (co)variance estimator in classic regression analysis, namely, the mean squared residual (MSR). Khuri (1996, p. 385) gives the following MSR estimator of $\sigma_{i;h}$:

$$\hat{\sigma}_{i;h}^N(\mathbf{d}, \boldsymbol{\omega}) = \frac{(\hat{\mathbf{f}}_i(\mathbf{d}, \boldsymbol{\omega}) - \hat{\mathbf{g}}_i(\mathbf{d}, \boldsymbol{\omega}))^T(\hat{\mathbf{f}}_h(\mathbf{d}, \boldsymbol{\omega}) - \hat{\mathbf{g}}_h(\mathbf{d}, \boldsymbol{\omega}))}{N - q} \quad \text{for } h = 0, \ldots, r-1, \tag{5}$$

where $\hat{\mathbf{g}}_i(\mathbf{d}, \boldsymbol{\omega}) = \mathbf{X}\hat{\boldsymbol{\beta}}_i^N(\mathbf{d}, \boldsymbol{\omega})$ denotes the $N \times 1$ vector with regression predictors for response i that follow from (2) and (3), and $n > q$, which implies a nonsaturated design such as a CCD.

In this paragraph, we summarize the results of convergence in probability and in distribution of the OLS estimators $\hat{\boldsymbol{\beta}}_i^N$ in (3) and the MSR estimators $\hat{\sigma}_{i;h}^N$ in (5); these results are standard in classic regression analysis, and they are required by the delta method and the generalized form of Wald's statistic. Suppose that the white-noise assumption holds per response type and the sequence $\{1/N(\mathbf{X}^T\mathbf{X})\}$ converges to a positive definite matrix as the sample size N increases (i.e., $N \to \infty$ such that $m_l \to \infty \; \forall l$). Then, Theil (1971, p. 362) proves that the sequence $\{\hat{\boldsymbol{\beta}}_i^N\}$ and the sequence $\{\hat{\sigma}_{i;h}^N\}$ converge in probability to $\boldsymbol{\beta}_i$ and $\sigma_{i;h}$, respectively, as a result of a law of large numbers. Furthermore, Theil (1971, p. 378) proves that—as a result of a central limit theorem—the sequence $\{\sqrt{N}(\hat{\boldsymbol{\beta}}_i^N - \boldsymbol{\beta}_i)\}$ converges in distribution to the multivariate normal with zero mean vector and covariance matrix $\boldsymbol{\Psi}_i$ with $\boldsymbol{\Psi}_i$ given by (4) replacing $\hat{\sigma}_{i;i}$ with $\sigma_{i;i}$.

2.1. Selecting the Size of the Local Experimental Area

Determining an appropriate size of the local experimental area requires a trade-off between bias and variance. When bias is of concern, it is recommended to locate input combinations close to the center of the local area—which may be explained by the Taylor expansion. When variance is of concern, the design points should be placed as far away from the local center as possible.

Some publications on this problem are Zazanis and Suri (1993), Safizadeh (2002), and Glasserman (2003, pp. 377–386). For finite-difference methods—which can be considered as one-at-a-time designs—Zazanis and Suri (1993), and Glasserman (2003) prove that the optimal size of the local experimental area depends on the second-order or third-order derivatives of the

unknown functions f_i in (1) and their unknown variances. Safizadeh's (2002) main finding is that both bias and variance of the gradient estimator decrease as the size of the local area decreases to zero but remains positive; his lower bound on the size of the local area is given by the signal/noise ratios

$$\hat{\gamma}^N_{i;\,t;\,t'} = \frac{\hat{\beta}^N_{i;\,t;\,t'}}{\sqrt{\widehat{\mathrm{var}}(\hat{\beta}^N_{i;\,t;\,t'})}} \quad \text{for } i = 0, \ldots, r-1,\ t,\ t' = 1, \ldots, k, \quad (6)$$

where $\widehat{\mathrm{var}}(\hat{\beta}^N_{i;\,t;\,t'})$ is the corresponding element on the main diagonal of the estimated covariance matrix for $\hat{\boldsymbol{\beta}}^N_i$ in (4). In (6), we consider the signal/noise ratios of two-factor interactions $\hat{\beta}^N_{i;\,t;\,t'}(t \neq t')$ and pure quadratic effects $\hat{\beta}^N_{i;\,t;\,t'}(t = t')$ only, because our stopping rule will be applied at points near the optimal solution where the quadratic effects may become significant.

To determine the size of the local experimental area, we proceed as follows. Suppose that $\mathbf{d}_1$ is a feasible point for (1); detecting the feasibility of $\mathbf{d}_1$ is explained in Stage 1 of §3.2. This $\mathbf{d}_1$ is considered to be the central input vector of the CCD. We also need to determine the remaining $n-1$ input vectors; in two-dimensional space with $\mathbf{d}_1 = (d_{1;\,1}, d_{1;\,2})^T$, a 2^2 full-factorial design is used as the R_V part of the CCD, so $n_V = 4$ and $n = 9$.

Step 1. We determine the factorial input vectors in the R_V part by changing the components of $\mathbf{d}_1$ by a user-defined percentage.

Step 2. We determine the positive and negative axial input vectors as follows. Suppose that we want to have all input vectors but $\mathbf{d}_1$ to lie on the surface of a hypersphere centered at $\mathbf{d}_1$. Then, the radius R of this hypersphere can be determined by the Euclidean distance of a factorial input vector to $\mathbf{d}_1$. Now, the $2k$ positive and negative axial input vectors are obtained by, respectively, adding R to and subtracting R from one of the components of $\mathbf{d}_1$ while keeping all other components fixed at those of $\mathbf{d}_1$.

Step 3. We simulate the n input vectors and estimate $\hat{\boldsymbol{\beta}}^N_i$, $\widehat{\boldsymbol{\Psi}}^N_i$, and $\hat{\gamma}^N_{i;\,t;\,t'}$ through (3), (4), and (6) respectively. If the estimated $\hat{\gamma}^N_{i;\,t;\,t'}$ are "acceptable," then we use the estimated $\hat{\boldsymbol{\beta}}^N_i$ and $\widehat{\boldsymbol{\Psi}}^N_i$ in our stopping rule in §3; otherwise, we reduce (e.g., halve) the percentage in Step 1 and repeat all computations in Steps 1–3.

Unfortunately, there are no general guidelines for determining an appropriate percentage in Step 1 prior to simulating the input vectors; hence the preceding three steps may be repeated more than once. Furthermore, in §4, we will also experiment with axial input vectors that do not lie on the surface of the hypersphere.

3. Heuristic Stopping Rule Based on the KKT Conditions

In this section, we introduce the stopping rule that tests the KKT conditions at a given input vector. Throughout the test we assume that the f_i are locally approximated by (2), with unknown coefficients estimated through (3); furthermore, the unknown (co)variances of the f_i are estimated through (5).

Our stopping rule consists of three stages, where the outcome of the first stage will be used in the second stage, and the outcome of the second stage will be used in the third stage. All three stages are performed at the center of the local CCD design, previously denoted by $\mathbf{d}_1$. Prespecified Type I error rates α_1, α_2, and α_3 are allocated to Stages 1, 2, and 3, respectively, such that $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$; because of Bonferroni's inequality, this makes α an upper bound for the Type I error probability of the whole procedure.

3.1. Analysis of the Three Stages of the Stopping Rule

3.1.1. Stage 1: Finding the Binding Constraints. The first stage is performed to obtain an index set $A(\mathbf{d}_1)$ that consists of the indices of the binding (active) constraints at $\mathbf{d}_1$. This is accomplished through a *Student t-test*, detailed in Stage 1 of §3.2, applied to each of the $r-1$ "slacks" of the $r-1$ constraints in (1):

$$\bar{s}_j(\mathbf{d}_1, \boldsymbol{\omega}_1) = \bar{f}_j(\mathbf{d}_1, \boldsymbol{\omega}_1) - a_j \quad j = 1, \ldots, r-1, \quad (7)$$

where

$$\bar{f}_j(\mathbf{d}_1, \boldsymbol{\omega}_1) = (f_j(\mathbf{d}_1, \boldsymbol{\omega}_{1_1}) + \cdots + f_j(\mathbf{d}_1, \boldsymbol{\omega}_{1_{m_1}}))/m_1, \quad (8)$$

and $\boldsymbol{\omega}_{1_{m_1}}$ is the seed used for the replicate m_1; a_j and f_j were defined in (1). These $r-1$ tests can result in either one of the following two outcomes for $A(\mathbf{d}_1)$:

1. We classify a constraint as binding whenever its observed slack value is not significantly positive or negative. If there is at least one such slack in the t-test, then there is at least one binding constraint at $\mathbf{d}_1$, which means that $A(\mathbf{d}_1)$ is not empty: $A(\mathbf{d}_1) \neq \varnothing$.

2. If each of the $r-1$ slacks in the t-test is either significantly positive or significantly negative, then $A(\mathbf{d}_1) = \varnothing$.

In the following two stages, we consider both cases, namely, the constrained optimum case with $A(\mathbf{d}_1) \neq \varnothing$ and the unconstrained optimum case with $A(\mathbf{d}_1) = \varnothing$; we explain the unconstrained case as a special case of the constrained case.

3.1.2. Stage 2: Checking the Linear Dependencies Among the Gradients of the Binding Constraints. Given the index set $A(\mathbf{d}_1)$, the second stage shows whether the gradients of the binding constraints at $\mathbf{d}_1$ are linearly independent. Let $\boldsymbol{\Gamma}$ denote the $k \times$

$|A(\mathbf{d}_1)|$ matrix, which has the gradients of the binding constraints at $\mathbf{d}_1$ as columns; $|\cdot|$ is the cardinality of the finite set $A(\mathbf{d}_1)$. We check the condition number of $\mathbf{\Gamma}$ as an indirect (but practical) indicator of the degree of linear dependence of its columns. The ideal value for the condition number is one, and the larger the condition number is, the more linearly dependent the columns are; see, for example, Gill et al. (2000, p. 29). Note that this stage is performed only when $|A(\mathbf{d}_1)| \geq 2$ because if there is a single binding constraint (hence, $\mathbf{\Gamma}$ has a single column), the condition number of $\mathbf{\Gamma}$ is always one. The condition number is checked through the *percentile method* (Efron and Tibshirani 1993, pp. 170–174), which uses a bootstrap procedure and order statistics. A step-by-step implementation of the method will be introduced in Stage 2 of §3.2.

The necessity of checking the condition number of $\mathbf{\Gamma}$ can be explained as follows. Let $\mathbf{d}^0$ be an unknown local minimizer of (1). Given a constraint qualification, the KKT conditions for $\mathbf{d}^0$ are

$$\begin{aligned} &\nabla E[f_0(\mathbf{d}^0, \boldsymbol{\omega})] - \sum_{j' \in A(\mathbf{d}^0)} \lambda_{j'}^0 \nabla E[f_{j'}(\mathbf{d}^0, \boldsymbol{\omega})] = \mathbf{0} \\ &\lambda_{j'}^0 \geq 0, \quad j' \in A(\mathbf{d}^0), \end{aligned} \tag{9}$$

where $\lambda_{j'}^0$ is the Lagrange multiplier for the j'th binding constraint at $\mathbf{d}^0$. The conditions in (9) imply that at $\mathbf{d}^0$, the gradient of the objective can be expressed as a nonnegative linear combination of the gradients of the binding constraints at $\mathbf{d}^0$. The constraint qualification is relevant when there are nonlinear constraints in (1); see Gill et al. (2000, p. 81). There are several forms of constraint qualification, many of which are of theoretical interest. A practical constraint qualification for nonlinear constraints is that the gradients of the binding constraints at $\mathbf{d}^0$ are linearly independent, which we assume for our problem in (1). Therefore, in this stage we statistically test the validity of our assumption.

If $\mathbf{\Gamma}$ is found to be well conditioned, then we proceed to the third stage. If $\mathbf{\Gamma}$ is ill conditioned, no action is taken; this issue is left for future research.

3.1.3. Stage 3: The Generalized Wald Statistic and the Delta Method. Given $A(\mathbf{d}_1)$ and a well-conditioned $\mathbf{\Gamma}$, Stage 3 is performed to determine whether $\mathbf{d}_1$ can be considered as an approximation of $\mathbf{d}^0$ by testing the KKT conditions at $\mathbf{d}_1$. We use the *delta method*, to show that under certain conditions, nonlinear statistics are asymptotically multivariate normally distributed, and the generalized form of *Wald's statistic* in Kodde and Palm (1986), to test composite hypotheses. The basic results and the technical conditions of both methods are introduced in this section; applications will be presented in Stage 3 of §3.2.

Based on the KKT conditions in (9), we test the following composite null hypothesis H_0 and its alternative hypothesis H_1:

$$\begin{aligned} H_0\colon\ & \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)] - \sum_{j' \in A(\mathbf{d}_1)} \lambda_{j'} \nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)] = \mathbf{0} \\ & \text{and} \quad \lambda_{j'} \geq 0 \quad \text{for all } j' \in A(\mathbf{d}_1), \\ H_1\colon\ & \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)] - \sum_{j' \in A(\mathbf{d}_1)} \lambda_{j'} \nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)] \neq \mathbf{0} \\ & \text{or} \quad \lambda_{j'} \ngeq 0 \text{ for some } j' \in A(\mathbf{d}_1). \end{aligned} \tag{10}$$

Because we do not know the functional forms of $E[f_i(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ explicitly, we use the regression metamodels in (2) within the local area. For a two-dimensional space where $\mathbf{d}_1 = (d_{1;1}, d_{1;2})^T$, this means that $\nabla E[f_i(\mathbf{d}_1, \boldsymbol{\omega}_1)] = (\beta_{i;1} + \beta_{i;1;2} d_{1;2} + 2\beta_{i;1;1} d_{1;1}, \beta_{i;2} + \beta_{i;1;2} d_{1;1} + 2\beta_{i;2;2} d_{1;2})^T$.

To use Wald's statistic in Kodde and Palm (1986), we obtain a slightly different form of (10), as follows. Suppose that $A(\mathbf{d}_1) \neq \varnothing$. It is well known in linear algebra that any $k \times 1$ vector can be uniquely decomposed into the sum of its range-space component and its null-space component; see, for example, Gill et al. (2000, p. 156). We apply this decomposition to $\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]$; i.e., $\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)] = \boldsymbol{\varepsilon} + \boldsymbol{\zeta}$, where $\boldsymbol{\varepsilon}$ is in the null space of $\mathbf{\Gamma}^T$ and $\boldsymbol{\zeta}$ is in the range space of $\mathbf{\Gamma}$. Now the residual $\boldsymbol{\varepsilon}$ is defined as

$$\begin{aligned} &\boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) \\ &\quad = [\mathbf{I}_k - \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T] \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \end{aligned} \tag{11}$$

where $\mathbf{I}_k - \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T$ is the orthogonal projection matrix that projects $\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ onto the null space of $\mathbf{\Gamma}^T$. Under H_0 in (10), $\boldsymbol{\varepsilon}$ is a zero vector.

It is also well known in linear algebra that any vector $\boldsymbol{\zeta}$ in the range space of $\mathbf{\Gamma}$ can be expressed as a linear combination of the columns of $\mathbf{\Gamma}$; i.e., $\boldsymbol{\zeta} = \mathbf{\Gamma}\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is an $|A(\mathbf{d}_1)| \times 1$ vector. Using $\mathbf{\Gamma}\boldsymbol{\lambda} = \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)] - \boldsymbol{\varepsilon}$ and (11), this $\boldsymbol{\lambda}$ is given by

$$\boldsymbol{\lambda}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) = (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]. \tag{12}$$

From (10) it follows that $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers $\lambda_{j'}$ because $\nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ for $j' \in A(\mathbf{d}_1)$ are the columns of $\mathbf{\Gamma}$, and $\mathbf{\Gamma}\boldsymbol{\lambda} = \sum_{j' \in A(\mathbf{d}_1)} \lambda_{j'} \nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)]$.

Now, (10) becomes

$$\begin{aligned} H_0\colon\ & \boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) = \mathbf{0} \quad \text{and} \\ & \boldsymbol{\lambda}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) \geq \mathbf{0}, \\ H_1\colon\ & \boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) \neq \mathbf{0} \quad \text{or} \\ & \boldsymbol{\lambda}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \mathbf{\Gamma}) \ngeq \mathbf{0}. \end{aligned} \tag{13}$$

In other words, (13) implies that at $\mathbf{d}_1$, the gradient $\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ of the objective function is within

the nonnegative cone generated by the gradients $\nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ of the binding constraints.

The following theorem and its corollary present the main results of this paper.

THEOREM 1. *Under H_0 in* (13), *the statistic* $\widehat{W}$,

$$\begin{aligned}\widehat{W} = {} & (\sqrt{N}\hat{\boldsymbol{\varepsilon}}_N)^T \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1} (\sqrt{N}\hat{\boldsymbol{\varepsilon}}_N) \\ & + (\sqrt{N}\hat{\boldsymbol{\lambda}}_N - \bar{\boldsymbol{\lambda}} - \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1} \sqrt{N}\hat{\boldsymbol{\varepsilon}}_N)^T \\ & \cdot (\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N} - \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N, \hat{\boldsymbol{\lambda}}_N})^{-1} \\ & \cdot (\sqrt{N}\hat{\boldsymbol{\lambda}}_N - \bar{\boldsymbol{\lambda}} - \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1} \sqrt{N}\hat{\boldsymbol{\varepsilon}}_N), \end{aligned} \tag{14}$$

has the following chi-bar-square distribution

$$\Pr\{\widehat{W} \geq u \mid \widehat{\boldsymbol{\Sigma}}, A(\mathbf{d}_1)\} = \sum_{c=0}^{|A(\mathbf{d}_1)|} w_c \Pr\{\chi^2_{k+|A(\mathbf{d}_1)|-c} \geq u\}, \tag{15}$$

where $\hat{\boldsymbol{\varepsilon}}_N$ *and* $\hat{\boldsymbol{\lambda}}_N$ *in* (14) *are consistent estimators of* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\lambda}$; $\bar{\boldsymbol{\lambda}}$ *is the orthogonal projection of* $\hat{\boldsymbol{\lambda}}_N$ *onto the feasible space defined by* $\boldsymbol{\varepsilon} = \mathbf{0}$ *and* $\boldsymbol{\lambda} \geq \mathbf{0}$; $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}$, $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N}$, $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N, \hat{\boldsymbol{\lambda}}_N}$, *and* $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N}$ *are the submatrices of* $\widehat{\boldsymbol{\Sigma}}$, *which is a consistent estimator of the covariance matrix of* $\sqrt{N}(\hat{\boldsymbol{\varepsilon}}_N^T - \boldsymbol{\varepsilon}^T, \hat{\boldsymbol{\lambda}}_N^T - \boldsymbol{\lambda}^T)^T$ *and*

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N} & \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N, \hat{\boldsymbol{\lambda}}_N} \\ \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} & \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N} \end{bmatrix}; \tag{16}$$

u in (15) *is the critical value; $\chi^2_{k+|A(\mathbf{d}_1)|-c}$ is the central chi-square random variable with $k + |A(\mathbf{d}_1)| - c$ degrees of freedom; and w_c is the weight denoting the probability that c of the $|A(\mathbf{d}_1)|$ components of $\bar{\boldsymbol{\lambda}}$ are positive.*

For a proof of the theorem, we refer to Kodde and Palm (1986).

We now comment on the weights w_c in (15). These w_c are such that $w_c \geq 0$ for each c, and $w_0 + \cdots + w_{|A(\mathbf{d}_1)|} = 1$. In general, the computation of these weights may be very complicated; see Shapiro (1988). In Stage 3 of §3.2, we will estimate the w_c through parametric bootstrapping (Efron and Tibshirani 1993).

Now, suppose that $A(\mathbf{d}_1) = \varnothing$. This simplifies (13) to

$$\begin{aligned} H_0: \ & \boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \boldsymbol{\Gamma}) = \mathbf{0}, \\ H_1: \ & \boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \boldsymbol{\Gamma}) \neq \mathbf{0}, \end{aligned} \tag{17}$$

where $\boldsymbol{\varepsilon}(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)], \boldsymbol{\Gamma}) = \nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]$.

COROLLARY 1. *Under H_0 in* (17), *the statistic* $\widehat{W}_{cl}$,

$$\widehat{W}_{cl} = (\sqrt{N}\hat{\boldsymbol{\varepsilon}}_N)^T \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1} (\sqrt{N}\hat{\boldsymbol{\varepsilon}}_N), \tag{18}$$

has the central chi-square distribution with k degrees of freedom

$$\Pr\{\widehat{W}_{cl} \geq u \mid \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}\} = \Pr\{\chi^2_k \geq u\}. \tag{19}$$

Note that $\widehat{W}_{cl}$ is the classic Wald statistic, and hence the proof of the corollary follows from, for example, Lehmann (1999, pp. 533–534). The test statistic in (18) can also be found in Shapiro and Homem-de-Mello (1998), who assume strict complementarity; i.e., $E_{\boldsymbol{\omega}}[\bar{s}_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)]\lambda_{j'} = 0$ implies $\lambda_{j'} > 0$ for each binding constraint j'.

In the remainder of this subsection, we present the critical issues and the results of the delta method and the generalized form of Wald's statistic.

Rubinstein and Shapiro (1993, p. 256) present the delta method as follows. Let $\{\boldsymbol{\gamma}_N\}$ be a sequence of random vectors converging in probability to a vector $\boldsymbol{\mu}$ (i.e., $\boldsymbol{\gamma}_N \xrightarrow{P} \boldsymbol{\mu}$). Let $\{\tau_N\}$ be a sequence of positive numbers tending to infinity such that $\tau_N(\boldsymbol{\gamma}_N - \boldsymbol{\mu})$ converges in distribution to a random vector $\boldsymbol{\delta}$. Let $\mathbf{g}$ be a mapping that is differentiable at $\boldsymbol{\mu}$ such that $\nabla\mathbf{g}(\boldsymbol{\mu}) \neq \mathbf{0}$, where $\nabla\mathbf{g}(\boldsymbol{\mu})$ denotes the Jacobian matrix of $\mathbf{g}$ at $\boldsymbol{\mu}$. Then, the delta method implies that $\tau_N(\mathbf{g}(\boldsymbol{\gamma}_N) - \mathbf{g}(\boldsymbol{\mu}))$ has the same limiting distribution as $(\nabla\mathbf{g}(\boldsymbol{\mu}))^T\boldsymbol{\delta}$.

The critical issues in the delta method are the existence of $\nabla\mathbf{g}(\boldsymbol{\mu}) \neq \mathbf{0}$ and $\{\boldsymbol{\gamma}_N\}$, which satisfies the two types of convergence mentioned above. For our application of the delta method, we prove that $\boldsymbol{\varepsilon}$ in (11) and $\boldsymbol{\lambda}$ in (12) are differentiable with respect to the gradient $\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ of the objective and the gradients $\nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)]$ of the binding constraints at $\mathbf{d}_1$; see the online supplement. Furthermore, because $\hat{\boldsymbol{\beta}}_i^N \xrightarrow{P} \boldsymbol{\beta}_i$, $\{\sqrt{N}(\hat{\boldsymbol{\beta}}_i^N - \boldsymbol{\beta}_i)\}$ converges in distribution to the multivariate normal, and the gradient estimators are simply given by linear combinations of the components of $\hat{\boldsymbol{\beta}}_i^N$ (see the two-dimensional example below (10)), the sequence of the gradient estimators converges in probability to the true gradient, and it converges in distribution to multivariate normal. Hence, with $\tau_N = \sqrt{N}$, our application of the delta method shows that $\sqrt{N}(\hat{\boldsymbol{\varepsilon}}_N^T - \boldsymbol{\varepsilon}^T, \hat{\boldsymbol{\lambda}}_N^T - \boldsymbol{\lambda}^T)^T$ is asymptotically multivariate normal with zero mean vector and covariance matrix $\widehat{\boldsymbol{\Sigma}}$ in (16).

Kodde and Palm (1986) present the generalized form of Wald's statistic, as follows. They consider H_0: $\boldsymbol{\gamma}_1 = \mathbf{0}$, $\boldsymbol{\gamma}_2 \geq \mathbf{0}$ and H_1: $\boldsymbol{\gamma}_1 \neq \mathbf{0}$, $\boldsymbol{\gamma}_2 \ngeq \mathbf{0}$, where both $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are vectors. Critical to their test is the existence of a consistent estimator $\hat{\boldsymbol{\gamma}}_N$ of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T)^T$ such that $\sqrt{N}(\hat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma})$ is asymptotically multivariate normal with zero mean vector and covariance matrix $\widehat{\boldsymbol{\Sigma}}$. Then, they obtain the generalized form of Wald's statistic by projecting $\hat{\boldsymbol{\gamma}}_N$ orthogonally onto the feasible space defined by $\boldsymbol{\gamma}_1 = \mathbf{0}$ and $\boldsymbol{\gamma}_2 \geq \mathbf{0}$.

In our application of the generalized Wald statistic, we have $\hat{\boldsymbol{\varepsilon}}_N$ and $\hat{\boldsymbol{\lambda}}_N$ as consistent estimators of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\lambda}$, whereas the delta method shows that $\sqrt{N}(\hat{\boldsymbol{\varepsilon}}_N^T - \boldsymbol{\varepsilon}^T, \hat{\boldsymbol{\lambda}}_N^T - \boldsymbol{\lambda}^T)^T$ is asymptotically multivariate

normal. Then, the generalized form of Wald's statistic $\widehat{W}$ in (14) is obtained by orthogonally projecting $(\hat{\boldsymbol{\varepsilon}}_N^T, \hat{\boldsymbol{\lambda}}_N^T)^T$ onto the feasible space defined by $\boldsymbol{\varepsilon} = \mathbf{0}$ and $\boldsymbol{\lambda} \geq \mathbf{0}$. Under the null hypothesis in (13), this $\widehat{W}$ has the mixture of chi-square distributions defined in (15).

3.2. Application of the Stopping Rule

3.2.1. Stage 1: Finding the Binding Constraints. To determine the index set $A(\mathbf{d}_1)$, we test the following null hypothesis and its alternative hypothesis per constraint j in (1):

$$\begin{aligned} H_{0,j}: \ E_{\boldsymbol{\omega}}[\bar{s}_j(\mathbf{d}_1, \boldsymbol{\omega}_1)] = 0 \quad (j = 1, \ldots, r-1), \\ H_{1,j}: \ E_{\boldsymbol{\omega}}[\bar{s}_j(\mathbf{d}_1, \boldsymbol{\omega}_1)] \neq 0, \end{aligned} \tag{20}$$

where the $\bar{s}_j$ were defined in (7).

We assume that the f_i have finite variances $\sigma_{i;i}$. Because of the asymptotic multivariate normality of $(\bar{f}_1, \ldots, \bar{f}_{r-1})^T$ defined in (8), the slack vector $(\bar{s}_1, \ldots, \bar{s}_{r-1})^T$ also has an asymptotic multivariate normal distribution. Under $H_{0,j}$ in (20), the following statistic has a Student t-distribution with $\nu = N - q$ degrees of freedom:

$$t_{j,\nu} = \frac{\bar{s}_j(\mathbf{d}_1, \boldsymbol{\omega}_1)}{\sqrt{\hat{\sigma}_{j;j}(\mathbf{d}_1, \boldsymbol{\omega}_1)/m_1}} \quad (j = 1, \ldots, r-1), \tag{21}$$

where $\hat{\sigma}_{j;j}$ follows from (5) with $i = h = j$. Because there are $r-1$ slacks to be tested, choosing $\alpha_1/(r-1)$ for the Type I error rate of each test in (21) results in an upper bound of α_1 for the Type I error probability of Stage 1. Let $t_{\alpha_1/\{2(r-1)\},\nu}$ denote the $1 - \alpha_1/\{2(r-1)\}$ quantile of t_ν; i.e., the critical value for the two-sided test of (20). We then reject $H_{0,j}$ if the absolute value of $t_{j,\nu}$ exceeds $t_{\alpha_1/\{2(r-1)\},\nu}$.

We assume that $\mathbf{d}_1$ is a feasible point for (1). It is possible to verify this assumption as follows. The $(r-1)$ estimated slacks $\bar{s}_j$ in (7) converge to their true values almost surely (a.s.) because of the strong law of large numbers. If these slacks are nonnegative ($\bar{s}_j \geq 0$), then $\mathbf{d}_1$ is feasible for (1) a.s. However, if $\mathbf{d}_1$ is found to be infeasible, then searching for a feasible point requires the determination of a search direction and a step size; we leave this problem for future research.

3.2.2. Stage 2: Checking the Linear Dependencies Among the Gradients of the Binding Constraints. We test

$$\begin{aligned} H_0: \ \boldsymbol{\Gamma} \text{ is well conditioned}, \\ H_1: \ \boldsymbol{\Gamma} \text{ is not well conditioned}, \end{aligned} \tag{22}$$

through the percentile method. We proceed through the following steps, where we use the standard notation for bootstrapped values, namely, the superscript asterisk.

Step 1. We sample the bootstrap values $\hat{\boldsymbol{\beta}}_{j'}^*$ for all binding constraints j' from their asymptotic multivariate normal distributions with mean vectors and covariance matrices given by the estimates in (3) and (4).

Step 2. We compute the condition number κ^* of an estimate of $\hat{\boldsymbol{\Gamma}}$, which is a consistent estimator of $\boldsymbol{\Gamma}$, from the bootstrapped values $\hat{\boldsymbol{\beta}}_{j'}^*$. (Remember that in a two-dimensional space, the columns of $\boldsymbol{\Gamma}$ are given by $(\beta_{j';1} + \beta_{j';1;2} d_{1;2} + 2\beta_{j';1;1} d_{1;1}, \beta_{j';2} + \beta_{j';1;2} d_{1;1} + 2\beta_{j';2;2} d_{1;2})^T$.)

Step 3. We repeat Steps 1 and 2 K times (K denotes the so-called "bootstrap sample size"); we choose $K = 1{,}000$.

Step 4. Let the corresponding order statistics be $\kappa_{(1)}^* < \cdots < \kappa_{(1,000)}^*$. The index y of the critical value (the upper confidence limit) is given by

$$y = \lceil Kp + z_{\alpha_2} \sqrt{Kp(1-p)} \rceil, \tag{23}$$

where z_{α_2} is the $1 - \alpha_2$ quantile of the standard normal distribution and $p = 0.5$ (corresponding with the median). We reject H_0 if the condition number of the current estimate of $\hat{\boldsymbol{\Gamma}}$ exceeds $\kappa_{(y)}^*$ because smaller values of the condition number are preferred.

3.2.3. Stage 3: The Generalized Wald Statistic and the Delta Method. To determine whether $\mathbf{d}_1$ may be considered as an approximation of $\mathbf{d}^0$, we test the null and the alternative hypotheses in (13) through the generalized Wald statistic in (14), which has the chi-bar-square distribution in (15). In the rest of this section, we give the necessary formulas for the application of our rule; these formulas are derived in the online supplement.

The submatrices $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}$, $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N}$, $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N, \hat{\boldsymbol{\lambda}}_N}$, and $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N}$ of $\widehat{\boldsymbol{\Sigma}}$ in (16) are given by

$$\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N} = (\nabla \hat{\boldsymbol{\varepsilon}}_N)^T \widehat{\boldsymbol{\Psi}}_N (\nabla \hat{\boldsymbol{\varepsilon}}_N), \qquad \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N} = (\nabla \hat{\boldsymbol{\lambda}}_N)^T \widehat{\boldsymbol{\Psi}}_N (\nabla \hat{\boldsymbol{\lambda}}_N),$$
$$\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N, \hat{\boldsymbol{\lambda}}_N} = (\nabla \hat{\boldsymbol{\varepsilon}}_N)^T \widehat{\boldsymbol{\Psi}}_N (\nabla \hat{\boldsymbol{\lambda}}_N),$$
$$\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} = (\nabla \hat{\boldsymbol{\lambda}}_N)^T \widehat{\boldsymbol{\Psi}}_N (\nabla \hat{\boldsymbol{\varepsilon}}_N),$$

where the $(k|A'(\mathbf{d}_1)|) \times k$ and $(k|A'(\mathbf{d}_1)|) \times |A(\mathbf{d}_1)|$ matrices $\nabla \hat{\boldsymbol{\varepsilon}}_N$ and $\nabla \hat{\boldsymbol{\lambda}}_N$ are consistent estimators of the Jacobian matrices of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\lambda}$, the $(k|A'(\mathbf{d}_1)|) \times (k|A'(\mathbf{d}_1)|)$ matrix $\widehat{\boldsymbol{\Psi}}_N$ is a consistent estimator of the covariance matrix of the estimated gradients of the objective and the binding constraints at $\mathbf{d}_1$, and $A'(\mathbf{d}_1)$ is the index set of all binding constraints at $\mathbf{d}_1$ plus the index of the objective function.

First, we introduce $\widehat{\boldsymbol{\Psi}}_N$; then we give formulas for $\nabla \hat{\boldsymbol{\varepsilon}}_N$ and $\nabla \hat{\boldsymbol{\lambda}}_N$. This $\widehat{\boldsymbol{\Psi}}_N$ is a symmetric matrix that can be partitioned:

$$\widehat{\boldsymbol{\Psi}}_N = \begin{bmatrix} \widehat{\boldsymbol{\Psi}}_{N;0;0} & \widehat{\boldsymbol{\Psi}}_{N;0;j'} \\ \widehat{\boldsymbol{\Psi}}_{N;0;j'}^T & \widehat{\boldsymbol{\Psi}}_{N;j';h'} \end{bmatrix}, \tag{24}$$

where $\widehat{\mathbf{\Psi}}_{N;0;0}$ is the $k \times k$ symmetric covariance submatrix of the estimated gradient of the objective with

$$N\Bigg[\widehat{\mathrm{var}}(\hat{\beta}_{0;t}) + \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'}^2 \widehat{\mathrm{var}}(\hat{\beta}_{0;t;t'}) + 4d_{1;t}^2 \widehat{\mathrm{var}}(\hat{\beta}_{0;t;t}) + 2\sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{0;t;t'}) + 4d_{1;t} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{0;t;t}) + 4d_{1;t} \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{0;t;t})\Bigg] \tag{25}$$

on the diagonal, and

$$\begin{aligned} N\Bigg[&\widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{0;v}) + \sum_{v'=1,\,v'\neq v}^{k} d_{1;v'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{0;v;v'}) + 2d_{1;v} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{0;v;v}) \\ &+ \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{0;v}) + \sum_{t'=1,\,t'\neq t}^{k} \sum_{v'=1,\,v'\neq v}^{k} d_{1;'} d_{1;v'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{0;v;v'}) \\ &+ 2d_{1;v} \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{0;v;v}) + 2d_{1;t} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{0;v}) \\ &+ 2d_{1;t} \sum_{v'=1,\,v'\neq v}^{k} d_{1;v'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{0;v;v'}) \\ &+ 4d_{1;t} d_{1;v} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{0;v;v})\Bigg], \end{aligned} \tag{26}$$

with $v = 1, \ldots, k, v \neq t$ off the main diagonal; all variances and covariances can be estimated from the corresponding components of (4) after replacing $\hat{\sigma}_{i;i}^N$ by $\hat{\sigma}_{0;0}^N$. In (24), $\widehat{\mathbf{\Psi}}_{N;0;j'}$ is the $k \times (k|A(\mathbf{d}_1)|)$ covariance submatrix of the estimated gradients of the objective and the binding constraints at $\mathbf{d}_1$, which has

$$\begin{aligned} N\Bigg[&\widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{j';p}) + \sum_{p'=1,\,p'\neq p}^{k} d_{1;p'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{j';p;p'}) + 2d_{1;p} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t}, \hat{\beta}_{j';p;p}) \\ &+ \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{j';p}) + \sum_{t'=1,\,t'\neq t}^{k} \sum_{p'=1,\,p'\neq p}^{k} d_{1;t'} d_{1;p'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{j';p;p'}) \\ &+ 2d_{1;p} \sum_{t'=1,\,t'\neq t}^{k} d_{1;t'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t'}, \hat{\beta}_{j';p;p}) + 2d_{1;t} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{j';p}) \\ &+ 2d_{1;t} \sum_{p'=1,\,p'\neq p}^{k} d_{1;p'} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{j';p;p'}) \\ &+ 4d_{1;t} d_{1;p} \widehat{\mathrm{cov}}(\hat{\beta}_{0;t;t}, \hat{\beta}_{j';p;p})\Bigg] \end{aligned} \tag{27}$$

as components for $p = 1, \ldots, k$. Again, all covariances can be estimated from (4) after replacing $\hat{\sigma}_{i;i}^N$ by $\hat{\sigma}_{0;j'}^N$. The $(k|A(\mathbf{d}_1)|) \times (k|A(\mathbf{d}_1)|)$ submatrix $\widehat{\mathbf{\Psi}}_{N;j';h'}$ in (24) consists of the covariance matrix of each estimated gradient of the binding constraint j' whose components can be estimated through (25) and (26) after replacing the subindex 0 by j' and its covariance matrices with other binding constraints h' (i.e., $h' \in A(\mathbf{d}_1)$), whose components can be estimated through (27) after replacing the subindex 0 by h'.

Now we give the formulas for $\nabla\hat{\boldsymbol{\varepsilon}}_N$ and $\nabla\hat{\boldsymbol{\lambda}}_N$. The first k rows of the $\nabla\hat{\boldsymbol{\varepsilon}}_N$ and $\nabla\hat{\boldsymbol{\lambda}}_N$ are consistent estimators of $\partial\boldsymbol{\varepsilon}^T/\partial(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)])$ and $\partial\boldsymbol{\lambda}^T/\partial(\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)])$:

$$(\mathbf{I}_k - \hat{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T)^T$$

and

$$((\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T)^T.$$

The remaining $k|A(\mathbf{d}_1)|$ rows of $\nabla\hat{\boldsymbol{\varepsilon}}_N$ and $\nabla\hat{\boldsymbol{\lambda}}_N$ are consistent estimators of $\partial\boldsymbol{\varepsilon}^T/\partial(\nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)])$ and $\partial\boldsymbol{\lambda}^T/\partial(\nabla E[f_{j'}(\mathbf{d}_1, \boldsymbol{\omega}_1)])$, respectively,

$$\begin{aligned} &(-[\mathbf{e}_{j'}^T(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]](\mathbf{I}_k - \hat{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T) \\ &\quad - \hat{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\mathbf{e}_{j'}\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]^T \\ &\quad \cdot (\mathbf{I}_k - \hat{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T))^T \end{aligned} \tag{28}$$

and

$$\begin{aligned} &(-[\mathbf{e}_{j'}^T(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]](\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T \\ &\quad + (\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\mathbf{e}_{j'}\nabla E[f_0(\mathbf{d}_1, \boldsymbol{\omega}_1)]^T(\mathbf{I}_k - \hat{\mathbf{\Gamma}}(\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T))^T, \end{aligned} \tag{29}$$

where $\mathbf{e}_{j'}$ is the $|A(\mathbf{d}_1)| \times 1$ vector of zeros, except a 1 in position j'. Obviously, (28) and (29) are estimated for every binding constraint j'.

To estimate the weights w_c in (15), we use the bootstrap procedure consisting of the following four steps.

Step 1. We sample the bootstrap values $(\hat{\boldsymbol{\varepsilon}}_N^*, \hat{\boldsymbol{\lambda}}_N^*)$ from their asymptotic multivariate normal distribution.

Step 2. We count the number of times that the bootstrapped values $\sqrt{N}\hat{\boldsymbol{\lambda}}_N^* - \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\lambda}}_N, \hat{\boldsymbol{\varepsilon}}_N} \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\varepsilon}}_N}^{-1}\sqrt{N}\hat{\boldsymbol{\varepsilon}}_N^*$ have exactly $0, 1, \ldots, |A(\mathbf{d}_1)|$ positive components; i.e., we update the counters $k_0, k_1, \ldots, k_{|A(\mathbf{d}_1)|}$.

Step 3. We repeat Steps 1 and 2 $K = 1{,}000$ times.

Step 4. We use the final fractions $k_0/K, k_1/K, \ldots, k_{|A(\mathbf{d}_1)|}/K$ to estimate the weights $w_0, w_1, \ldots, w_{|A(\mathbf{d}_1)|}$.

In the application of our rule, we estimate $\widehat{W}$ in (14) and use this estimate to obtain the p-values for every $\chi^2_{k+|A(\mathbf{d}_1)|-c}$ in (15). The final p-value of the test is the weighted sum of all p-values. H_0 in (13) is rejected if this final p-value is smaller than the Type I error rate, which is α_3. Similarly, after estimating $\widehat{W}_{cl}$ in (18) and obtaining the corresponding p-value, H_0 in (17) is rejected if the p-value is smaller than α_3.

4. Numerical Examples

In the following two subsections, we estimate the Type I and Type II error rates of the test in (15). The rates of the test in (19) for the unconstrained problem are given in the online supplement. We consider a simple "toy" problem and the (s, S) inventory problem with a service-level constraint. In the third subsection, we present conclusions from these numerical experiments.

In the toy problem, we consider quadratic functions that are to be approximated by the regression metamodels in (2); hence, our assumption of the validity of the metamodels (i.e., $\mu_{\epsilon_i} = 0$) is satisfied. Then, we first consider the case where the variance is constant within the local experimental area (i.e., homogeneous variances), and later we relax this assumption (i.e., heterogeneous variances) to examine its effects on the estimated Type I and Type II error rates. We emphasize that this toy problem provides controlled laboratory settings. In the (s, S) inventory problem, however, neither the assumption of valid metamodels nor the assumption of homogeneous variance is satisfied (i.e., the white-noise assumption is not satisfied). Therefore, this problem lays the groundwork for evaluating the performance of our test in realistic discrete-event simulation studies, which is the main intent of this paper.

In the following two subsections, the prespecified upper bound for the Type I error rate of the whole (three-stage) procedure is $\alpha = 10\%$. To measure the variability in the performance of our test, we obtain 2,000 macro replicates for the toy problem and 1,000 macro replicates for the inventory problem.

We implement the experiments on a PC with Windows XP, Intel Pentium 4 CPU of 1.60 GHz, and 1.00 GB RAM. The computer program is coded in MATLAB 7.6. Two thousand macro replicates of the toy problem take about five CPU minutes, whereas 1,000 macro replicates of the inventory problem take about three CPU hours.

4.1. Toy Problem

We consider the following simple example:

$$\begin{aligned} \text{minimize} \quad & E[(d_1+8)^2 + 5(d_2+8)^2 + \epsilon_0] \\ \text{subject to} \quad & E[-(d_1-3)^2 - d_2^2 - d_1 d_2 + \epsilon_1] \geq -4, \\ & E[-d_1^2 - 3(d_2+1.061)^2 + \epsilon_2] \geq -9, \end{aligned} \tag{30}$$

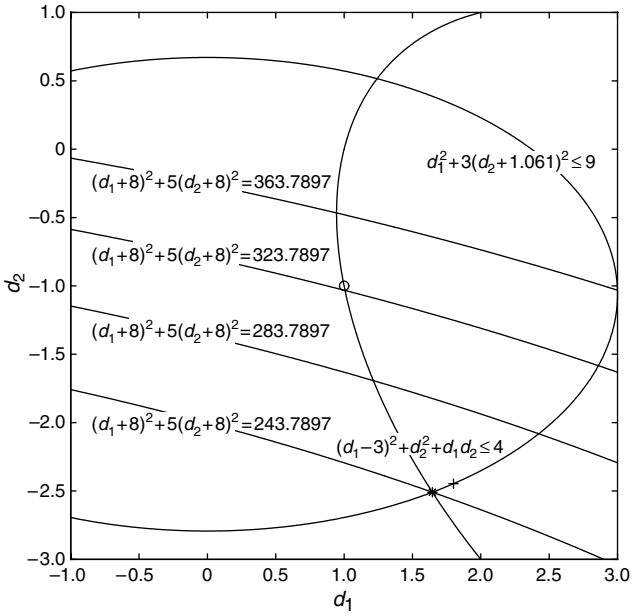

Figure 1 Toy Example Defined in (30)

where $(\epsilon_0, \epsilon_1, \epsilon_2)^T$ is multivariate normal with zero mean vector and covariance matrices given later in §§4.1.1 and 4.1.2. Figure 1 shows that the optimal input vector d_0 is $(1.6458, -2.5091)^T$; substitution into (30) gives an optimal mean objective value of 243.7897.

We experiment with three different local areas with the central points $(1.6458, -2.5091)^T$, $(1, -1)^T$, and $(1.8, -2.4466)^T$; see the points $*$, $\circ$, and $+$ in Figure 1. We consider the following cases: (i) 5% and 1.25% as the percentages to determine the local area size, and (ii) R and $2R$ to determine the axial points for each percentage. Furthermore, after obtaining Figures 1 and 2 in the online supplement, which show how the estimated Type I error rates of the KKT test change with respect to the number of replicates m_l, we decide to use (iii) $m_l = 250$ and $m_l = 850$ for all cases and for every input combination $l (l = 1, \ldots, n)$.

We use a CCD design, which consists of a 2^2 full-factorial design augmented with a central point and four axial points. For the central point $(1.6458, -2.5091)^T$ and the 5% percentage, we find the other eight local design points as follows. We first determine input vectors 1 through 4 in Table 1 by changing the coordinates of $(1.6458, -2.5091)^T$ by 2.5% (so the range is 5%). Then, we determine the radius $R = 0.075$, and obtain input vectors 6 through 9 by adding R to and subtracting R from one of the coordinates of $(1.6458, -2.5091)^T$. In the last two columns of Table 1, the axial points are obtained by adding and subtracting $2R$; hence, the input vectors do not lie on the sphere centered at $(1.6458, -2.5091)^T$.

Table 1 **CCD with the Central Point** $(1.6458, -2.5091)^T$

	Axial points given by R		Axial points given by $2R$	
Input vector	d_1	d_2	d_1	d_2
1	1.6869	−2.4464	1.6869	−2.4464
2	1.6869	−2.5718	1.6869	−2.5718
3	1.6047	−2.4464	1.6047	−2.4464
4	1.6047	−2.5718	1.6047	−2.5718
5	1.6458	−2.5091	1.6458	−2.5091
6	1.7208	−2.5091	1.7958	−2.5091
7	1.6458	−2.4341	1.6458	−2.3591
8	1.5708	−2.5091	1.4958	−2.5091
9	1.6458	−2.5841	1.6458	−2.6591

The allocation of the Type I error rates to each stage is as follows. At $(1.6458, -2.5091)^T$, both constraints are binding, so we apply all three stages of the test; the Type I error rates are $\alpha_1 = 3\%$, $\alpha_2 = 3\%$, and $\alpha_3 = 4\%$. At $(1, -1)^T$ and $(1.8, -2.4466)^T$, however, there is a single binding constraint. Hence, Stage 2 of the test is skipped at these points, and the Type I error rates are $\alpha_1 = 5\%$ and $\alpha_3 = 5\%$.

4.1.1. Homogeneous Variance Case. For the local area with the central point $(1.6458, -2.5091)^T$, we select $\sigma_{0,0} = 25$, $\sigma_{1,1} = 9$, and $\sigma_{2,2} = 16$. For the local area around $(1, -1)^T$, we select $\sigma_{0,0} = 12$, $\sigma_{1,1} = 8$, and $\sigma_{2,2} = 9$. For the local area around $(1.8, -2.4466)^T$, we select $\sigma_{0,0} = 10$, $\sigma_{1,1} = 7$, and $\sigma_{2,2} = 4$. For all three local areas, we let ρ denote the correlation between the responses and select $\rho_{0,1} = -0.2$, $\rho_{0,2} = 0.7$, and $\rho_{1,2} = -0.4$. We assume constant variance only locally, so as the local area changes, the variance changes as well.

In the online supplement, we give the average signal/noise ratios over 2,000 macro replicates and their sample standard deviations estimated through 2,000 macro replicates for the local areas centered at $(1.6458, -2.5091)^T$, $(1, -1)^T$, and $(1.8, -2.4466)^T$. Furthermore, Tables 2–4 summarize the results of the three stages of our test. We offer the following comments:

1. Increasing the number of replicates from 250 to 850 results in more accurate estimates of the signal/noise ratios, which give perfect results for all three stages of the test; see Tables 2–4. However, even with 250 replicates, we obtain good results for all three stages, with larger local areas (i.e., axial $= 2R$) giving slightly better results than smaller ones (i.e., axial $= R$).

2. Given Figure 1, we know (but our procedure does not) that both constraints in (30) are active at the true optimum. Hence, the number 1,945 (in the second row and second column of Table 2) is the number of macro replicates out of 2,000 macro replicates in which Stage 1 detects that both constraints are active at the optimum through the t-statistic in (21). Hence, the estimated Type I error rate is $1 - 1{,}945/2{,}000 = 0.0275$, which is lower than the prespecified $\alpha_1 = 0.03$.

Table 2 **Estimated Type I Error Rates for the Local Area with the True Optimal Point as the Central Point** $(1.6458, -2.5091)^T$ **and Locally Homogeneous Variances**

Local area	Both constraints active	$\boldsymbol{\Gamma}$ found ill-conditioned	KKT rejected
Size = 5%, rep. = 250 axial = R	1,945/2,000 = 0.9725	10/1,945 = 0.0051	43/1,935 = 0.0222
Size = 5%, rep. = 850 axial = R	1,952/2,000 = 0.9760	0/1,952 = 0	2/1,952 = 0.0010
Size = 5%, rep. = 250 axial = $2R$	1,936/2,000 = 0.9680	0/1,936 = 0	7/1,936 = 0.0036
Size = 5%, rep. = 850 axial = $2R$	1,952/2,000 = 0.9760	0/1,952 = 0	0/1,952 = 0

3. The number 10 (second row, third column of Table 2) is the number of macro replicates out of these 1,945 macro replicates in which H_0 in (22)—stating that $\boldsymbol{\Gamma}$ is well conditioned—is rejected through the test in (23). The fraction $10/1{,}945 = 0.0051$ estimates the conditional Type I error rate (conditional on the event that the binding set is detected correctly); this is lower than the nominal $\alpha_2 = 0.03$. The unconditional estimate of the Type I error rate, namely, $10/2{,}000 = 0.005$, is also lower than $\alpha_2 = 0.03$.

4. The number 43 (second row, fourth column of Table 2) is the number of macro replicates out of these 1,935 macro replicates in which H_0 in (13)—stating that the KKT conditions hold—is rejected through Wald's generalized test in (15). The fraction $43/1{,}935 = 0.0222$, which estimates the conditional Type I error rate (conditional on the events that the binding set is detected correctly and $\boldsymbol{\Gamma}$ is well conditioned), is lower than the nominal $\alpha_3 = 0.04$. The

Table 3 **Estimated Power for the Local Area Farthest Away from the True Optimum Centered at** $(1, -1)^T$ **and Locally Homogeneous Variances**

Local area	Constraint 1 active	KKT rejected
Size = 5%, rep. = 250 axial = R	1,953/2,000 = 0.9765	1,949/1,953 = 0.9980
Size = 5%, rep. = 850 axial = R	1,952/2,000 = 0.9760	1,952/1,952 = 1
Size = 5%, rep. = 250 axial = $2R$	1,953/2,000 = 0.9765	1,947/1,953 = 0.9969
Size = 5%, rep. = 850 axial = $2R$	1,952/2,000 = 0.9760	1,952/1,952 = 1

Table 4 Estimated Power for the Intermediate Local Area Centered at $(1.8, -2.4466)^T$ and Locally Homogeneous Variances

Local area	Constraint 2 active	KKT rejected
Size = 5%, rep. = 250 axial = R	1,962/2,000 = 0.9810	1,897/1,962 = 0.9669
Size = 5%, rep. = 850 axial = R	1,941/2,000 = 0.9705	1,941/1,941 = 1
Size = 5%, rep. = 250 axial = $2R$	1,962/2,000 = 0.9810	1,952/1,962 = 0.9949
Size = 5%, rep. = 850 axial = $2R$	1,941/2,000 = 0.9705	1,941/1,941 = 1

Table 5 Estimated Type I Error Rates for the Local Area Centered at $(1.6458, -2.5091)^T$ and Locally Heterogeneous Variances

Local area	Both constraints active	$\boldsymbol{\Gamma}$ found ill-conditioned	KKT rejected
Size = 5%, rep. = 250 axial = R	1,945/2,000 = 0.9725	0/1,945 = 0	65/1,945 = 0.0334
Size = 5%, rep. = 850 axial = R	1,946/2,000 = 0.9730	0/1,946 = 0	29/1,946 = 0.0149
Size = 5%, rep. = 250 axial = $2R$	1,945/2,000 = 0.9725	0/1,945 = 0	59/1,945 = 0.0303
Size = 5%, rep. = 850 axial = $2R$	1,946/2,000 = 0.9730	0/1,946 = 0	11/1,946 = 0.0057

unconditional estimate of the Type I error rate is $43/2{,}000 = 0.0215$, which is also lower than $\alpha_3 = 0.04$.

5. Table 2 shows that increasing the number of replicates from 250 to 850 does improve the estimates of the Type I error rates for (15); this increase in replicates decreases the conditional estimated Type I error rates from 0.0222 to 0.0010 (axial = R) and from 0.0036 to 0 (axial = $2R$), and the unconditional ones from $43/2{,}000 = 0.0215$ to 0.0010 (axial = R) and from 0.0035 to 0 (axial = $2R$).

6. In Table 3, the conditional estimates of Type II error rates are $4/1{,}953 = 0.0021$ (the number of replicates is 250, or rep. = 250, axial = R), $6/1{,}953 = 0.0031$ (rep. = 250, axial = $2R$), and 0 (rep. = 850, both axial = R and $2R$); see the last column. In Table 4, these estimates are $65/1{,}962 = 0.0331$ (rep. = 250, axial = R), $10/1{,}962 = 0.0051$ (rep. = 250, axial = $2R$), and 0 (rep. = 850, both axial = R and $2R$), which are slightly worse than those in Table 3; see again the last column. These results are especially surprising for $(1.8, -2.4466)^T$, which is very close to the true optimal solution $(1.6458, -2.5091)^T$; see Figure 1. In Table 3, the unconditional estimates of Type II error rates are $4/2{,}000 = 0.002$ (rep. = 250, axial = R), $6/2{,}000 = 0.003$ (rep. = 250, axial = $2R$), and 0 (rep. = 850, both axial = R and $2R$), and in Table 4, they are given by $65/2{,}000 = 0.0325$ (rep. = 250, axial = R), $10/2{,}000 = 0.005$ (rep. = 250, axial = $2R$), and 0 (rep. = 850, both axial = R and $2R$).

The average signal/noise ratios are at acceptable levels; however, to compare the results of the test, we also experiment with 2.5% and 1.25% as the user-defined percentages. We do not show the results for 2.5%, which gives only slightly worse results than those in Tables 2–4. We present the results of 1.25% in the online supplement, where now in Stage 2, $\boldsymbol{\Gamma}$ is found ill conditioned in 584 out of 1,955 macro replicates. Hence, the estimated conditional Type I error rate is $584/1{,}955 = 0.2987$, which exceeds the nominal $\alpha_2 = 0.03$. Increasing the number of replicates to 850, however, this estimate decreases to $59/1{,}960 = 0.0301$; see the online supplement.

4.1.2. Heterogeneous Variance Case. Now we assume that the variance changes with the input combination l as follows, where $l = 1, \ldots, 9$ and max denotes the maximum: $\sigma_{0,0} = 2(\max\{d_{l;1}, d_{l;2}\})^2$, $\sigma_{1,1} = 1.5(\max\{d_{l;1}, d_{l;2}\})^2$, and $\sigma_{2,2} = 2.5(\max\{d_{l;1}, d_{l;2}\})^2$. In all three local areas, we use the same correlations, namely, $\rho_{0,1} = -0.2$, $\rho_{0,2} = 0.7$, and $\rho_{1,2} = -0.4$. We start experimenting with the 5% percentage, because this gives the best results in the homogeneous case. This percentage gives acceptable average signal/noise ratios, so we do not use a smaller percentage. We summarize our results in Tables 5–7 and give the following comments:

1. Relaxing the constant-variance assumption has no impact on the estimated Type I error rates of the t-test; see the second columns of the Tables 5–7.

2. Heterogeneous variances do not affect the conditional Type II error rate estimates of the generalized Wald test; see the last columns of Tables 6 and 7.

3. The conditional estimates of the Type I error rates of the generalized Wald test, however, are slightly worsened by heterogeneous variances; see the last column of Table 5.

Table 6 Estimated Power for the Local Area Centered at $(1, -1)^T$ and Locally Heterogeneous Variances

Local area	Constraint 1 active	KKT rejected
Size = 5%, rep. = 250 axial = R	1,959/2,000 = 0.9795	1,959/1,959 = 1
Size = 5%, rep. = 850 axial = R	1,946/2,000 = 0.9730	1,946/1,946 = 1
Size = 5%, rep. = 250 axial = $2R$	1,959/2,000 = 0.9795	1,959/1,959 = 1
Size = 5%, rep. = 850 axial = $2R$	1,946/2,000 = 0.9730	1,946/1,946 = 1

Table 7 **Estimated Power for the Local Area Centered at $(1.8, -2.4466)^T$ and Locally Heterogeneous Variances**

Local area	Constraint 2 active	KKT rejected
Size = 5%, rep. = 250 axial = R	1,949/2,000 = 0.9745	1,900/1,949 = 0.9749
Size = 5%, rep. = 850 axial = R	1,948/2,000 = 0.9740	1,945/1,948 = 0.9985
Size = 5%, rep. = 250 axial = $2R$	1,949/2,000 = 0.9745	1,923/1,949 = 0.9867
Size = 5%, rep. = 850 axial = $2R$	1,948/2,000 = 0.9740	1,942/1,948 = 0.9969

4.2. (s, S) Inventory Problem with a Service-Level Constraint

Bashyam and Fu (1998) assume an infinite-horizon, periodic review inventory system with continuous-valued and IID demands and full backlogging of orders. In each period, orders are received at the beginning of the period, the demand for the period is subtracted, and order review occurs at the end of the period. An order is placed when the inventory position (stock on hand plus outstanding suppliers' orders minus customers' backorders) falls below the reorder level s; the order amount is the difference between the order-up-to level S and the current inventory position. Suppliers' orders can cross in time, which makes an analytical solution impossible.

The objective function $E_{\boldsymbol{\omega}}[f_0(\mathbf{d}, \boldsymbol{\omega})]$ in (1) is the steady-state expected total costs, namely, the sum of order setup, ordering, and holding costs. There is a single stochastic constraint $E_{\boldsymbol{\omega}}[f_1(\mathbf{d}, \boldsymbol{\omega})]$, which is the expected steady-state "fill-rate," i.e., the fraction of demand directly met from stock on hand. We select a target fill-rate of 0.90; thus $a_j = 0.9$ in (1).

Demands are exponentially distributed with mean 100; order lead times are Poisson distributed with mean 6. The setup cost is 36 per order, the order cost is 2 per unit, and the holding cost is 1 per period per unit. Each simulation run lasts for 2,500 periods. After examining Figures 3 and 4 in the online supplement—showing how the estimated Type I error rates change with respect to the number of replicates m_l—and considering that 1,000 macro replicates with $m_l = 2$ take about three CPU hours, we decide to make $m_l = 2$ replicates per input combination. Our simulation of the inventory system starts with the inventory position and the inventory level (stock on hand) at S without any outstanding suppliers' orders.

Bashyam and Fu (1998) do not report their estimated optimal (s^*, S^*). Kleijnen and Wan (2007) find by brute-force simulation experiments that the best estimate of (s^*, S^*) is $s = 1{,}020$ and $S = 1{,}075$—which has the average cost of 623.70 with a standard error of 2.36 and the average fill-rate of 0.899 with a standard error of 0.004.

We again use a CCD design with nine input vectors, where now the central point is $(1{,}020, 1{,}075)^T$. Now we start experimenting with 2.5% and obtain acceptable average signal/noise ratios. However, we also halve this percentage and experiment with 1.25%; the average signal/noise ratios corresponding to both 2.5% and 1.25% are given in the online supplement. Our brute-force experiments show that the expected fill-rate is binding at $(1{,}020, 1{,}075)^T$. Because there is only one binding constraint, Stage 2 of the test is skipped, and the nominal Type I error rates are $\alpha_1 = 5\%$ and $\alpha_3 = 5\%$. We summarize our results for both 2.5% and 1.25% in Table 8.

1. The estimated Type I error rate of the t-test $(1 - 948/1{,}000 = 0.052)$ exceeds the nominal $\alpha_1 = 0.05$; this difference is insignificant when tested through the binomial distribution approximated by the normal distribution with mean 0.05 and variance $0.05(1 - 0.05)/1{,}000$ for all relevant (less than or equal to 10%) nominal Type I error probabilities.

2. The estimated conditional Type I error rate of the generalized Wald test exceeds the nominal $\alpha_3 = 0.05$ for all three local areas, where the best result $(1 - 896/953 = 0.0598)$ is given by the smallest area (size = 1.25%, axial = R) and the worst result $(1 - 675/935 = 0.2781)$ is given by the largest area (size = 2.5%, axial = $2R$). Remember that in the toy problem, the larger local areas result in better test results than the smaller ones. This contradiction is to be expected, because in the toy problem the true functions are quadratic polynomials, so there is no bias in the OLS estimators in (3) when we approximate these functions through regression metamodels in (2). For the (s, S) inventory problem, however, the OLS estimators may be biased because the two simulation responses are not quadratic polynomials, and the bias increases with the size of the local area (Taylor expansion argument).

Table 8 (s, S) **Inventory Problem: Estimated Type I Error Rates for the Local Area Centered at the True Optimum** $(1{,}020, 1{,}075)^T$

Local area	Fill-rate constraint active	KKT rejected
Size = 2.5%, rep. = 2 axial = R, run = 2,500 periods	948/1,000 = 0.948	147/948 = 0.1551
Size = 2.5%, rep. = 2 axial = $2R$, run = 2,500 periods	935/1,000 = 0.935	260/935 = 0.2781
Size = 1.25%, rep. = 2 axial = R, run = 2,500 periods	953/1,000 = 0.953	57/953 = 0.0598
Size = 1.25%, rep. = 2 axial = $2R$, run = 2,500 periods	940/1,000 = 0.940	109/940 = 0.1159

4.3. Conclusions About Numerical Examples

We summarize our results of the numerical examples as follows. In general, the estimated (un)conditional Type I error rate exceeds its nominal value; this suggests that in case our stopping rule is implemented in an optimization algorithm, the rule will cause the optimization program to run longer. On the other hand, the estimated (un)conditional Type II error rate of our rule suggests that it is very difficult for a suboptimal point to be considered as an estimated optimal solution; i.e., our rule tends to reject KKT conditions even at the "true" optimal solution. This issue needs to be further investigated.

5. Conclusions and Future Research

In this paper, we derive an asymptotic heuristic stopping rule that combines three statistical tests for the KKT first-order necessary optimality conditions for constrained or unconstrained simulation-optimization problems. Our rule uses linear regression, the bootstrap-percentile method, the delta method, and the generalized Wald statistic. We estimate the performance of our rule through extensive simulation experiments with a toy problem and an inventory problem. Our numerical results are encouraging; i.e., the alternative hypothesis (the KKT conditions do not hold) is often rejected at suboptimal points. The empirical Type I error rate (at the true optimum), however, exceeds its nominal value.

Future research may incorporate our new stopping rule within an optimization algorithm. Moreover, more realistic discrete-event simulation models need to be investigated.

Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://joc.journal.informs.org/.

Acknowledgments

The authors thank Alexander Shapiro (University of Georgia Tech) for his personal communication suggesting the delta method to prove that the residual has a multivariate normal distribution. Moreover, they thank Bertrand Melenberg (Tilburg University) for several technical comments on a previous version of their paper. Finally, the authors thank two anonymous reviewers and the associate editor for their valuable comments on two earlier versions, which led to the present paper.

References

Bashyam, S., M. C. Fu. 1998. Optimization of (s, S) inventory systems with random lead times and a service level constraint. *Management Sci.* **44**(12, Part 2) 243–256.

Bettonvil, B., E. del Castillo, J. P. C. Kleijnen. 2009. Statistical testing of optimality conditions in multiresponse simulation-based optimization. *Eur. J. Oper. Res.* **199**(2) 448–458.

Efron, B., R. J. Tibshirani. 1993. *An Introduction to the Bootstrapped*. Chapman & Hall, New York.

Fu, M. C. 2002. Optimization for simulation: Theory vs. practice. *INFORMS J. Comput.* **14**(3) 192–215.

Gill, P. E., W. Murray, M. H. Wright. 2000. *Practical Optimization*, 12th ed. Academic Press, London.

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering*. Springer, New York.

Higle, J. L., S. Sen. 1991. Statistical verification of optimality conditions for stochastic programs with recourse. *Ann. Oper. Res.* **30**(1–4) 215–240.

Karaesmen, I., G. van Ryzin. 2004. Overbooking with substitutable inventory classes. *Oper. Res.* **52**(1) 83–104.

Khuri, A. I. 1996. Multiresponse surface methodology. S. Ghosh, C. R. Rao, eds. *Handbook of Statistics, Design and Analysis of Experiments*, Vol. 13. Elsevier, Amsterdam, 377–406.

Kleijnen, J. P. C. 2008. *Design and Analysis of Simulation Experiments*. Springer Science + Business Media, New York. [Chinese translation published by Publishing House of Electronics Industry, Beijing, 2010.]

Kleijnen, J. P. C., J. Wan. 2007. Optimization of simulated systems: OptQuest and alternatives. *Simulation Modelling Practice Theory* **15**(3) 354–362.

Kodde, D. A., F. C. Palm. 1986. Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* **54**(5) 1243–1248.

Lehmann, E. L. 1999. *Elements of Large-Sample Theory*. Springer, New York.

Rao, C. R. 1967. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. L. M. Lecam, J. Neyman, eds. *Proc. Fifth Berkeley Sympos. Math. Statist. Probab.*, Vol. I. University of California Press, Berkeley, 355–372.

Rubinstein, R. Y., A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, Chichester, UK.

Safizadeh, M. H. 2002. Minimizing the bias and variance of the gradient estimate in RSM simulation studies. *Eur. J. Oper. Res.* **136**(1) 121–135.

Shapiro, A. 1988. Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.* **56**(1) 49–62.

Shapiro, A., T. Homem-de-Mello. 1998. A simulation-based approach to two-stage stochastic programming with recourse. *Math. Programming* **81**(3) 301–325.

Spall, J. C. 2003. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, Hoboken, NJ.

Theil, H. 1971. *Principles of Econometrics*. John Wiley & Sons, New York.

Zazanis, M. A., R. Suri. 1993. Convergence rates of finite-difference sensitivity estimates for stochastic systems. *Oper. Res.* **41**(4) 694–703.